AWS Machine Learning Engineer Udacity Capstone Project: Proposal

Bashir Suleiman Muhammad
(Dated: January 30, 2021)

DOMAIN BACKGROUND

Traditional Drug Discovery process takes up to a decade for a single drug to reach the market, but with the emergence of new powerful technologies like AI, Drug Discovery process will become faster, easier and more accurate. More drugs could be produced within a short period of time.

Machine Learning Models play an important role when leveraging AI for Drug Discovery. Machine Learning Models could be used in all stages of Drug Discovery and generate accurate results. Machine Learning Models could be used in Target Validation, prediction of solubility of a drug and its activity towards a target protein among others.

Solubility, measures how well a solute dissolve in a particular solvent to give a homogeneous solution. It is one of the most important factors to consider during drug discovery among other Pharmokinetic properties of the drug. It will surely be a great concern for a patient to absorb a poorly soluble drug, because that could lead to other types of diseases like kidney stones, Diuresis among others.

"Low aqueous solubility is the major problem encountered with formulation development of new chemical entities as well as for the generic development. More than 40% NCEs (new chemical entities) developed in pharmaceutical industry are practically insoluble in water" [1]

PROBLEM STATEMENT

In this project, we are going to train a Machine Learning Model to predict the solubility of compounds directly from their Chemical Structures encoded in SMILE string using its four Molecular Descriptors which includes;

a. ESOL predicted log solubility in mols per litre
b. Minimum Degree
c. Molecular Weight
d. Number of H-Bond Donors
e. Number of Rings
f. Number of Rotatable Bonds
g. Polar Surface Area

The problem to be solved however is: *Predict the solubility of a given compound based on its Molecular Descriptors.*

DATASETS AND INPUTS

Delaney.csv is a **Standard Regression Dataset** gotten from kaggle [3]. It contains chemical structures and water solubility data for 1128 compounds. The data is already cleaned, processed and ready for use. However, we intend to use only highly ranked features (it's molecular descriptors) as inputs. Click here to access the datab

SOLUTION STATEMENT

'"*Linear regression is a linear model, e.g., a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x)*"[2].

In this case, since we have relationship between our multiple input variables(x), i.e the values of our molecular descriptors, and our output variable (y), i.e the predicted solubility in **mol/L**, we are going to use a **Linear Regression Model**.

BENCHMARK MODEL

We are going to train a Linear Regression Model and then we are going to tune it by changing various hyperparameters. We are then going to compare our results with the model containing the default hyperparameters to know which hyperparameters give the best results.

EVALUATION METRICS

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

RMSE calculates the average of how much our model's predictions deviates from the actual value while R2 calculates how well the model is trained to give accurate results from its inputs
We are going to calculate RMSE, R2 for each model to see how well it is trained.

PROJECT DESIGN

These are the steps I will take inorder to solve the proposed problem:

A) **Data Gathering**: The data was already gathered from Kaggle

B) **Data Pre-processing**: The dataset was already processed but will be checked again.

C) **Data Splitting**: The dataset will then be splitted **80%** for training and **20%** for testing

D) **Model Training**: I will train a **Linear Regression Model**, finetune it using set of hyperparameters and then compare their respective **R2 scores and RMSE scores** to determine the best set of hyperparameters for the model.

E) **Model Inference**: I will deploy the model to a AWS Sage maker's managed Endpoint that will allow me to make inferences from new data

References

[1] Savjani KT, Gajjar AK, Savjani JK. Drug solubility: importance and enhancement techniques. *ISRN Pharm*. 2012;2012:195727. doi:10.5402/2012/195727

[2] https://www.sciencedirect.com/topics/mathematics/simple-linear-regression-model

[3] https://www.kaggle.com/c/drug-solubility-challenge/data